Volume 10, Issue 3 March 2017

Bioinformatics up to Date

(Bioinformatics Infrastructure Facility, Biotechnology Division)
North-East Institute of Science & Technology
Jorhat - 785 006, Assam

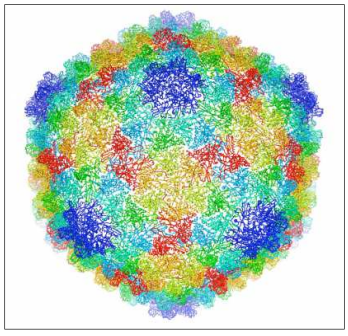
Inside…..

About us

The Bioinformatics Infrastructure Facility (BIF) at Biotechnology division, CSIR NEIST, Jorhat runs under the Biotechnology Information System Network (BTISnet) programme of DBT, Ministry of Science & Technology, and Government of India. The Centre was established on 2nd February, 2008 to promote innovation in Biological research and education through Bioinformatics accomplishment. The main goal is to facilitate and expose students and researchers from different academic institutions of North East India in Bioinformatics. The center conduct training and workshops for enlightening the use of bioinformatics applications in biological research and development. The Centre has access to global information through 24 hour high speed internet facility, and also e-journal facilities with DeLCON, Science Direct etc. To date the Centre has profoundly extended support in R & D work with a great intensity to different biological discipline including medicinal chemistry, computer aided drug design, genomics and proteomic data analysis etc.

cryoSPARC: Newly developed algorithms for rapid unsupervised cryo-EM structure determination

A new set of machine learning algorithms developed by U of T researchers that can generate 3D structures of tiny protein molecules may revolutionize the development of drug therapies for a range of diseases, from Alzheimer's to cancer.



The work published in journal *Nature Methods (February 06, 2017).*

This new set of algorithms reconstructs 3D structures of protein molecules using microscopic images. Since proteins are tiny even smaller than a wavelength of light - they can't be seen directly without using sophisticated techniques like electron cryomicroscopy (cryo-EM). This new method, CryoSPARC, is revolutionizing the way scientists can discover 3D protein structures, allowing the study of many proteins that simply could not be studied in the past. CryoSPARC™ is an easy to use software tool that enables rapid, unbiased structure discovery of proteins and molecular complexes from cryo-EM data.

The algorithms, which were co-developed by Fleet's former Post-Doctoral Fellow Marcus Brubaker, now an Assistant Professor at York University, could significantly aid in the development of new drugs because they provide a faster, more efficient means at arriving at the correct protein structure.

*[Source: **cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.** Ali Punjani et al. Nature Methods (February 06, 2017]*

SC3: New Tool to Analysis of Single-Cell RNA Data in Pre-Malignant Tumours

Scientists from Wellcome Trust Sanger Institute and their collaborators have developed a new analysis tool called Single Cell Consensus Clustering (SC3) that was able to show, which genes were expressed by individual cells in different genetic versions of a benign blood cancer. The work published in Nature Methods March 27. Single cell RNA sequencing can define cell types by revealing differences in the proteins produced by individual cells, however analyzing the data remains challenging. The new open source computer tool was shown to be more accurate and robust than existing methods of analysing single-cell RNA sequence data, and is freely available for researchers to use.

Recent advances in single-cell genomics technology has made it possible to separate individual cells from different tissues and organs, and measure the sets of RNA messages - called the transcriptome - which help give each cell its own identity. These individual transcriptomes can be used to define cell types and to understand the functions of healthy and diseased cells in the human body. This technology has enormous potential for biological research.

"The SC3 tool was able to use patterns of gene expression to distinguish, within an individual cancer, subclones that carried different mutations. This approach will help us define the cellular heterogeneity within each cancer, an important step towards improving cancer treatment," says Professor Tony Green, an author from the Wellcome Trust-MRC Stem Cell Institute and Cambridge University .

[***Source: SC3: consensus clustering of single-cell RNA-seq data.*** *Vladimir Yu Kiselev et al. Nature Methods (2017)*]

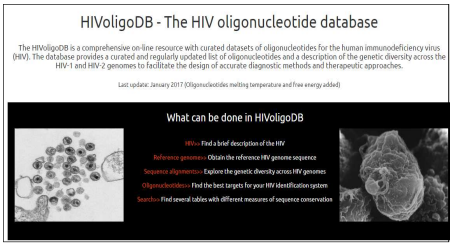
PCPPI: database for prediction of Penicillium–crop protein–protein interactions

Penicillium expansum, the causal agent of blue mold, is one of the most prevalent post-harvest pathogens, infecting a wide range of crops after harvest. In response, crops have evolved various defense systems to protect themselves against this and other pathogens. Penicillium–crop interaction is a multifaceted process and mediated by pathogen- and host-derived proteins. Identification and characterization of the inter-species protein–protein interactions (PPIs) are fundamental to elucidating the molecular mechanisms underlying infection processes between P. expansum and plant crops. Here, we have developed PCPPI, the Penicillium-Crop Protein–Protein Interactions database, which is constructed based on the experimentally determined orthologous interactions in pathogen–plant systems and available domain–domain interactions (DDIs) in each PPI. Thus far, it stores information on 9911 proteins, 439 904 interactions and seven host species, including apple, kiwifruit, maize, pear, rice, strawberry and tomato. Further analysis through the gene ontology (GO) annotation indicated that proteins with more interacting partners tend to execute the essential function. Significantly, semantic statistics of the GO terms also provided strong support for the accuracy of our predicted interactions in PCPPI. We believe that all the PCPPI datasets are helpful to facilitate the study of pathogen-crop interactions and freely available to the research community.

[Source: Database (Oxford) (2017) 2017 (1): baw170. DOI: https://doi.org/10.1093/database/baw170]

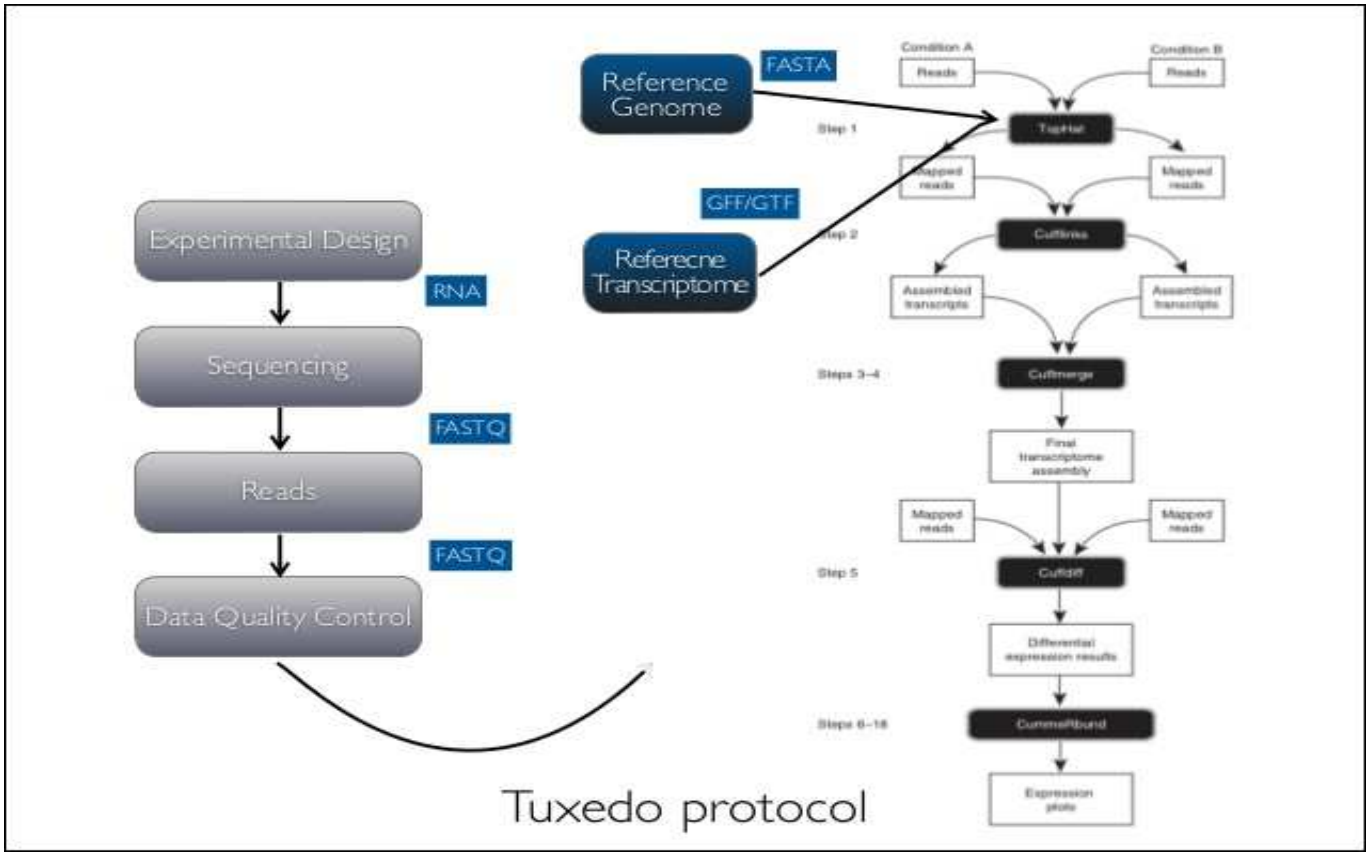
The HIV oligonucleotide database

The human immunodeficiency virus (HIV) is associated with one of the most widespread infectious disease, the acquired immunodeficiency syndrome (AIDS). The development of antiretroviral drugs and methods for virus detection requires a comprehensive analysis of the HIV genomic diversity, particularly in the binding sites of oligonucleotides. HIVoligoDB a versatile online database with oligonucleotides selected for the diagnosis of HIV and treatment of AIDS. Currently, the database provides an interface for visualization, analysis and download of 380 HIV-1 and 65 HIV-2 oligonucleotides annotated according to curated reference genomes. The database also allows the selection of the most conserved HIV genomic regions for the development of molecular diagnostic assays and sequence-based candidate therapeutics.



The HIVoligoDB (http://portugene.com/HIVoligoDB) includes 380 HIV -1 (236 primers and 144 probes) and 65 HIV-2 oligonucleotides (57 primers and 8 probes) retrieved from 54 peer-reviewed publications and the NCBI Probe Database. Each oligonucleotide has a specific database code (for example, HIV1ID0001) and is associated with a reference genome (K03455.1 for HIV-1 and M15390 for HIV-2). The database provides descriptive webpages (e.g. type of target, related publications) for each oligonucleotide and a search engine to access dynamic tables with numeric data and multiple sequence alignments with complete HIV genomes. The multiple sequence alignments were retrieved from the Los Alamos National Laboratory.

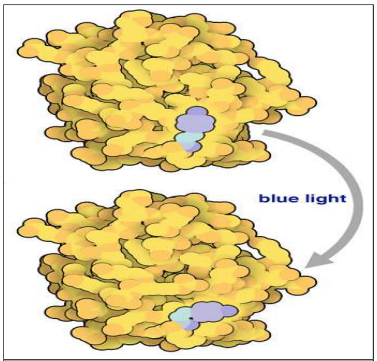
[Source: Database (Oxford) (2017) 2017 (1): bax005; https://doi.org/10.1093/database/bax005]



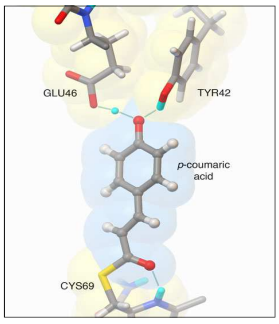
Molecule of Month

Photoactive Yellow Protein

PYP is a small, soluble protein found in purple sulfur bacteria, where it senses blue light. PYP is a perfect subject for study, since it forms large, stable crystals that diffract well, allowing us to get a detailed view of what is happening at the atomic level. The photocycle of PYP involves many separate, sequential steps that begin very quickly in femtoseconds, and end in about a second. A light-absorbing chromophore in the protein absorbs blue light and switches from a nearly straight (*trans*) conformation to a bent (*cis*) conformation in less than a picosecond. This conformational change is then thought to activate other sensing proteins in the bacterium, which control the direction that the bacterium is swimming.



This structure of photoactive yellow protein, solved by time-resolved Laue crystallography, includes a 50-50 mixture of the ground state (top) and the light-activated state (bottom).



The X-ray crystallographic structures of PYP have shown that the chromophore (*p*-coumaric acid, also known as 4'-hydroxycinnamic acid) is held tightly in the protein. It is covalently attached to a cysteine at one end, and it forms unusually short hydrogen bonds with two amino acids at the other end. To look at these hydrogen bonds more closely, researchers have used both X-ray diffraction and neutron diffraction to look at different aspects. X-ray crystallography sees the location of electrons in a structure, so it is useful for determining the location of heavier atoms, such as carbon, nitrogen and oxygen atoms.

[Source: http://pdb101.rcsb.org/motm/207]

Patents

Bioinformatics data processing systems

WO 2016148650 A1

Inventors : Davide Verzotto, Niranjan Nagarajan

ABSTRACT

Disclosed is a computer-implemented method of determining at least one optimal alignment of at least part of a first map to at least part of a second map or a plurality of second maps, wherein the maps are physical genome maps and/or restriction maps. The method comprises: receiving first map data indicative of a first ordered list of distances between features of the first map, receiving second map data indicative of a second ordered list of distances between features of the second map or second maps; generating, from the second map data, seed data indicative of a plurality of seeds, each seed comprising at least one of the distances in the second ordered list, wherein the features are restriction sites and distances are fragment sizes. The said method further comprises generating a plurality of candidate alignments from the seed data by searching at least part of the first ordered list to find at least approximate matches for respective seeds, and extending the approximate matches by dynamic programming; determining respective alignment scores for respective candidate alignments; and selecting one or more of the candidate alignments as an optimal alignment or optimal alignments, based on the alignment scores.

Kindly send us your feedback to

Dr Ratul Saikia
BIF Center, Biotechnology Group, BSTD
CSIR-North East Institute of Science and Technology, Jorhat, Assam
E-mail: rsaikia19@gmail.com